Shiv Hari Tewari / Int.J.Data.Sci. and Big Data Anal. 1(1) (2021) 63-79
https://doi.org/10.51483/IJDSBDA.1.1.2021.63-79 ISSN: 2710-2599

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https: / / www.svedbergopen.com /

Review Article Open Access

Necessity of data science for enhanced Cybersecurity

Shiv Hari Tewari[1*]

[1]M.Tech. (CSE), Birla Institute of Technology, Mesra, Ranchi, India. E-mail: tewarishivhari999@gmail.com

Article Info

Volume 1, Issue 1, February 2021
Received : 13 November 2020
Accepted : 11 January 2021
Published : 05 February 2021
doi: 10.51483/IJDSBDA.1.1.2021.63-79

Abstract

In a computing context, cybersecurity is undergoing massive shifts in technology and its operations in recent days, and data science is driving the change. Extracting security incident patterns or insights from cybersecurity data and building corresponding data-driven model, is the key to make a security system automated and intelligent. To understand and analyze the actual phenomena with data, various scientific methods, machine learning techniques, processes, and systems are used, which is commonly known as data science. In this paper, I have briefly described the data science its evolution its applications in cloud security and how cybersecurity data science came in existence what kind of advantages are given by Cybersecurity Data Science (CSDS) and its steps like, where the data is being gathered from relevant cybersecurity sources, and the analytics complement the latest data-driven patterns for providing more effective security solutions. The concept of cybersecurity data science allows making the computing process more actionable and intelligent as compared to traditional ones in the domain of cybersecurity. After that I have described the various upcoming challenges that can emerge after the frequent applications of CSDS, how machine learning and deep learning are applicable in it and types of algorithms that can be applicable in it. So, the overall paper is not only focuses on the origins of Data Science but it also describes its modern uses for the relevant cybersecurity field and data driven intelligent decision making system can protect our system from known and unknown cyber attacks.

Keywords: *Bitcoin, cryptocurrencies, extreme gradient boosting, deep neural networks.*

1. Introduction

In the era of the data-driven industries when industries are becoming more and more data centric, the data and information is the new treasure for any cyber attacker, attackers today don't focus on the transactions the industries has done but their more focus is on the information shared between two persons and so that they can attack again on the systems if they want, thus, the data security is the major concern for the today's major industries and data science can be the answer to this question as the industries today are focusing on securing their data so much, the data has became the new gold for the attackers, most of the attackers use the old methods to attack on any server of industries the most usual way to attack the systems are like phishing, injecting a malware, inserting the virus, DDoS or DoS attacks and man in middle attacks, and we can train our systems to detect such attacks even if the firewall or cyber guard of the system

* *Corresponding author: Shiv Hari Tewari, M.Tech. (CSE), Birla Institute of Technology, Mesra, Ranchi, India.*
E-mail: tewarishivhari999@gmail.com

compromises they can let us know the moment of compromise and even prevent the system from further attacks. The branch of data science which deals with the cyber security and its issues is known as the CSDS. The CSDS can be the answer to all the unknown attacks happening to the servers and can provide the desired level of security to the industries. Today, when AI based malwares are coming in the light, now industries need something much powerful than the antimalware, antivirus and firewall and data science can work in such way to provide this level of security to the systems and the industries. The CSDS basically focus on training the systems with many types of attacks happened already and uses proper machine learning algorithms to make systems learn how to detect the attack and if attack somehow attack happened then to prevent the files and information of the system from getting accessed by the attackers. Hence this is how the CSDS can bring the revolutionary change to the domain of the cyber security and it can even detect the AI based viruses and malwares which usually makes the antimalware and firewalls compromise easily and they are really difficult to find out in the systems as they usually leaves no traces behind in such kind of attacks CSDS can be a better option than many cyber security and antimalware software as we need a proper dataset and machine learning algorithm to train the system properly and remaining tasks can be done by the system by itself. Hence that's how CSDS can become the new guard for the industries today.

2. Literature review

The term "data science" was first coined in 2008 by D.J. Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market. Today, we will examine how the meaning of this job has evolved in history. In the recent years the data scientist has became a buzzword for the industries, since many industries require the data scientists today, it became a fashion for the people who belongs to IT to be the Data Scientists, many researchers has contributed their works in the field of data science. Since industries are becoming the data driven their reliability on data can pay them the cost in the form of cyber security, today, all the cyber attackers want is the access to the data of the particular industry and to know the vulnerabilities of the server of an organization and hence, *the data became the new gold for the cyber attackers.*

There are many researchers who has given their noble contribution for the applications of data science in enhancement of the cyber security, recently Sarkar et al. (2020a) has given their contribution to show how data science is applicable in the enhancement of the cyber security and can provide the industries a better and secured system with the help of machine learning algorithms and models. Earlier Sumeet Dua and Xian Du (2011) has proposed advanced data mining and machine learning in their data mining and machine learning in cybersecurity book. There are many other researchers who has proposed the models for the cybersecurity data science and the way of using to prevent the systems from cyber attacks.

Faheem and Muhammad (2018) has provided a systematic review for the architectural tactics for big data cybersecurity analytics system. Devakunchari and Sourabh (2019) presented the article about the applications of machine learning for cybersecurity.

Manjeet and Raymond (2018) discussed in their paper how the machine learning can be used to create an impenetrable cyber defense system by using the methods of machine learning like supervise, unsupervised and semi-supervised machine learning models.

Azzah et al. (2020) has presented the methods of artificial intelligence and its applications for providing cybersecurity and discussed various ways to prevent our systems from the cyber attacks and to overcome their vulnerabilities.

Thuraisingham et al. (2018) has presented the developments in applying Data science for cyber security and cyber security for data science and then discuss the applications in Social Media. There are many related works in data science for providing cybersecurity to the systems of the industries and make them little more robust and impenetrable than the usual. Hence these are the few related works on the cybersecurity data science and the applications of the data science in providing the cybersecurity and to prevent the systems from cyber attackers.

3. Data science - A brief history

First before moving to the point of security we should know about how data science came in the existence, why it became so much important for the modern technology and IT infrastructures. The story how data scientists became trendy is mostly the story of coupling of the mature discipline of statistics with a very young one computer science. The term "Data Science" has emerged only recently to specifically designate a new profession that is expected to make sense of the vast stores of the big data. But making sense of data has a long history and has been discussed by scientists, statisticians, librarians, computer scientists and others for others for years. The following timeline traces the

evolution of the term "Data Science" and its uses, attempts to define it and its related terms (https://www.oreilly.com/library/view/doing-data-science/9781449363871/ch01.html[origins of data science]).

In 1962 John W Tukey written in the "*The Future of Data Analysis*": For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and doubt. I have come to feel that my central interest is in *data analysis* Data analysis, and the parts of statistics which adhere to it, must take on the characteristics of science rather than those of mathematics data analysis is intrinsically an empirical science. How vital and how important. Is the rise of the stored-program electronic computer? In many instances the answer may surprise many by being 'important but not vital,' although in others there is no doubt but what the computer has been 'vital'. In 1947, Tukey coined the term "bit" which Claude Shannon used in his 1948 paper, "A Mathematical Theory of Communications." In 1977, Tukey published "*Exploratory Data Analysis"*, arguing that more emphasis needed to be placed on using data to suggest hypotheses to test and that Exploratory Data Analysis and Confirmatory Data Analysis "can and should proceed side by side."

In 1974, Peter Naur publishes "*Concise Survey of Computer Methods"* in Sweden and the United States. The book is a survey of contemporary data processing methods that are used in a wide range of applications. It is organized around the concept of data as defined in the *IFIP Guide to Concepts and Terms in Data Processing*: "[Data is] a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process". The preface to the book tells the reader that a course plan was presented at the IFIP Congress in 1968, titled, "Datalogy, the science of data and of data processes and its place in education", "and that in the text of the book", the term 'data science' has been used freely." Naur offers the following definition of data science: "The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."

In 1977,The International Association for Statistical Computing (IASC) is established as a Section of the ISI. "It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."

In 1989, Gregory Piatetsky-Shapiro organizes and chairs the first Knowledge Discovery in Databases (KDD) workshop. In 1995, it became the annual ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).

In September 1994, *Business Week* publishes a cover story on "Database Marketing": "Companies are collecting mountains of information about you, crunching it to predict how likely you are to buy a product, and using that knowledge to craft a marketing message precisely calibrated to get you to do so. An earlier flush of enthusiasm prompted by the spread of checkout scanners in the 1980s ended in widespread disappointment: Many companies were too overwhelmed by the sheer quantity of data to do anything useful with the information. Still, many companies believe they have no choice but to brave the database-marketing frontier."

In 1996, Members of the *International Federation of Classification Societies (IFCS)* meet in Kobe, Japan, for their biennial conference. For the first time, the term "data science" is included in the title of the conference ("Data science, classification, and related methods"). The IFCS was founded in 1985 by six country—and language-specific classification societies, one of which, *The Classification Society*, was founded in 1964. The classification societies have variously used the terms data analysis, data mining, and data science in their publications.

In 1997, the journal Data Mining and Knowledge Discovery is launched; the reversal of the order of the two terms in its title reflecting the ascendance of "data mining" as the more popular way to designate "extracting information from large databases."

In December 1999, Jacob Zahavi is quoted in "Mining Data for Nuggets of Knowledge" in Knowledge@Wharton: "Conventional statistical methods work well with small data sets. Today's databases, however, can involve millions of rows and scores of columns of data Scalability is a huge issue in data mining. Another technical challenge is developing models that can do a better job analyzing data, detecting non-linear relationships and interaction between elements. Special data mining tools may have to be developed to address web-site decisions."

In 2001, William S Cleveland publishes "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." It is a plan "to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called 'data science.'" Cleveland puts the proposed new discipline in the context of computer science and the contemporary work in data mining: "the benefit to the data analyst has been limited, because the knowledge among computer scientists about how to think of and approach the analysis of data is limited, just as the knowledge of computing environments by statisticians is limited. A merger of

knowledge bases would produce a powerful force for innovation. This suggests that statisticians should look to computing for knowledge today just as data science looked to mathematics in the past. Departments of data science should contain faculty members who devote their careers to advances in computing with data and who form partnership with computer scientists."

In the same year, Leo Breiman publishes "Statistical Modeling: The Two Cultures" (PDF): "There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools."

In April 2002, Launch of *Data Science Journal*, publishing papers on "the management of data and databases in Science and Technology. The scope of the journal includes descriptions of data systems, their publication on the internet, applications and legal issues." The journal is published by the Committee on Data for Science and Technology (CODATA) of the International Council for Science (ICSU).

In January 2003, Launch of *Journal of Data Science*: "By 'Data Science' we mean almost everything that has something to do with data: Collecting, analyzing, modeling yet the most important part is its applications all sorts of applications. This journal is devoted to applications of statistical methods at large. The *Journal of Data Science* will provide a platform for all data workers to present their views and exchange ideas."

In May 2005, Thomas H Davenport, Don Cohen, and Al Jacobson publish "Competing on Analytics," a Babson College Working Knowledge Research Center report, describing "the emergence of a new form of competition based on the extensive use of analytics, data, and fact-based decision making instead of competing on traditional factors, companies are beginning to employ statistical and quantitative analysis and predictive modeling as primary elements of competition. " The research is later published by Davenport in the *Harvard Business Review* (January 2006) and is expanded (with Jeanne G. Harris) into the book *Competing on Analytics: The New Science of Winning* (March 2007).

In September 2005, The National Science Board publishes "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century." One of the recommendations of the report reads: "The NSF, working in partnership with collection managers and the community at large, should act to develop and mature the career path for data scientists and to ensure that the research enterprise includes a sufficient number of high-quality data scientists." The report defines data scientists as "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection."

In January 2007, The Research Center for Dataology and Data Science is established at Fudan University, Shanghai, China. In 2009, two of the center's researchers, Yangyong Zhu and Yun Xiong, publish "Introduction to Dataology and Data Science," in which they state "Different from natural science and social science, Dataology and Data Science takes data in cyberspace as its research object. It is a new science." The center holds annual symposiums on Dataology and Data Science.

In July 2008, The JISC publishes the final report of a study it commissioned to "examine and make recommendations on the role and career development of data scientists and the associated supply of specialist data curation skills to the research community." The study's final report, "The Skills, Role & Career Structure of Data Scientists & Curators: Assessment of Current Practice & Future Needs," defines data scientists as "people who work where the research is carried out—or, in the case of data centre personnel, in close collaboration with the creators of the data—and may be involved in creative enquiry and analysis, enabling others to work with digital data, and developments in data base technology."

In January 2009, *Harnessing the Power of Digital Data for Science and Society* is published. This report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council states that "The nation needs to identify and promote the emergence of new disciplines and specialists' expert in addressing the complex and dynamic challenges of digital preservation, sustained access, reuse and repurposing of data. Many disciplines are seeing the emergence of a new type of data science and management expert, accomplished in the computer, information, and data sciences arenas and in another domain science. These individuals are key to the

current and future success of the scientific enterprise. However, these individuals often receive little recognition for their contributions and have limited career paths."

In January 2009, Hal Varian, Google's Chief Economist, tells the *McKinsey Quarterly*: "I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data, to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it, that's going to be a hugely important skill in the next decades. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it. I do think those skills—of being able to access, understand, and communicate the insights you get from data analysis—are going to be extremely important. Managers need to be able to access and understand the data themselves."

In March 2009, Kirk D Borne and other astrophysicists submit to the Astro2010 Decadal Survey a paper titled, "The Revolution in Astronomy Education: Data Science for the Masses" (PDF): "Training the next generation in the fine art of deriving intelligent understanding from data is needed for the success of sciences, communities, projects, agencies, businesses, and economies. This is true for both specialists (scientists) and non-specialists (everyone else: the public, educators and students, workforce). Specialists must learn and apply new data science research techniques in order to advance our understanding of the Universe. Non-specialists require information literacy skills as productive members of the 21st century workforce, integrating foundational skills for lifelong learning in a world increasingly dominated by data."

In May 2009, Mike Driscoll writes in "The Three Sexy Skills of Data Geeks": "with the Age of Data upon us, those who can model, munge, and visually communicate data—call us statisticians or data geeks—are a hot commodity." [Driscoll will follow up with The Seven Secrets of Successful Data Scientists in August 2010].

In June 2009, Nathan Yau writes in "Rise of the Data Scientist": "As we've all read by now, Google's chief economist Hal Varian commented in January that the next nice job in the next 10 years would be statisticians. Obviously, I whole-heartedly agree. I'd go a step further and say they're good now mentally and physically. However, if you went on to read the rest of Varian's interview, you'd know that by statisticians, he actually meant it as a general title for someone who is able to extract information from large datasets and then present something of use to non-data experts [Ben] Fry argues for an entirely new field that combines the skills and talents from often disjoint areas of expertise [computer science; mathematics, statistics, and data mining; graphic design, infovis and human-computer interaction]. And after two years of highlighting visualization on Flowing Data, it seems collaborations between the fields are growing more common, but more importantly, computational information design edges closer to reality. We're seeing *data scientists* people who can do it all emerge from the rest of the pack."

In the same year Troy Sadkowsky creates the data scientists group on LinkedIn as a companion to his website, datasceintists.com (which later became datascientists.net).

In February 2010, Kenneth Cukier writes in *The Economist* Special Report Data, Data Everywhere: a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/ artist to extract the nuggets of gold hidden under mountains of data.

In June 2010, Mike Loukides writes in "What is Data Science?": Data scientists combine entrepreneurship with patience, the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution. They are inherently interdisciplinary. They can tackle all aspects of a problem, from initial data collection and data conditioning to drawing conclusions. They can think outside the box to come up with new ways to view the problem, or to work with very broadly defined problems: 'here's a lot of data, what can you make from it?'

In September 2010, Drew Conway writes in "The Data Science Venn Diagram": "one need to learn a lot as they aspire to become a fully competent data scientist. Unfortunately, simply enumerating texts and tutorials does not untangle the knots. Therefore, in an effort to simplify the discussion, and add my own thoughts to what are already a crowded market of ideas, I present the Data Science Venn Diagram hacking skills, math and stats knowledge, and substantive expertise."

In May 2011, Pete Warden writes in "Why the term 'data science' is flawed but useful": "There is no widely accepted boundary for what's inside and outside of data science's scope. Is it just a faddish rebranding of statistics? I don't think so, but I also don't have a full definition. I believe that the recent abundance of data has sparked something new in the world, and when I look around I see people with shared characteristics who don't fit into traditional categories. These people tend to work beyond the narrow specialties that dominate the corporate and institutional world, handling everything from finding the data, processing it at scale, visualizing it and writing it up as a story. They also seem to start by looking at what the data can tell them, and then picking interesting threads to follow, rather than the traditional scientist's approach of choosing the problem first and then finding data to shed light on it."

In the same year and same month, David Smith writes in "'Data Science': What's in a name?": "The terms 'Data Science' and 'Data Scientist' have only been in common usage for a little over a year, but they've really taken off since then: many companies are now hiring for 'data scientists', and entire conferences are run under the name of 'data science'. But despite the widespread adoption, some have resisted the change from the more traditional terms like 'statistician' or 'quant' or 'data analyst' I think 'Data Science' better describes what we actually do: a combination of computer hacking, data analysis, and problem solving."

In June 2011, Matthew J Graham talks at the Astrostatistics and Data Mining in Large Astronomical Databases workshop about "The Art of Data Science" (PDF). He says: "To flourish in the new data-intensive environment of 21st century science, we need to evolve new skills… We need to understand what rules [data] obey, how it is symbolized and communicated and what its relationship to physical space and time is."

In September 2011, D.J. Patil writes in "Building Data Science Teams": "Starting in 2008, Jeff Hammerbacher (@hackingdata) and I sat down to share our experiences building the data and analytics groups at Facebook and LinkedIn. In many ways, that meeting was the start of data science as a distinct professional specialization, we realized that as our organizations grew, we both had to figure out what to call the people on our teams. 'Business analyst' seemed too limiting. 'Data analyst' was a contender, but we felt that title might limit what people could do. After all, many of the people on our teams had deep engineering expertise. 'Research scientist' was a reasonable job title used by companies like Sun, HP, Xerox, Yahoo, and IBM. However, we felt that most research scientists worked on projects that were futuristic and abstract, and the work was done in labs that were isolated from the product development teams. It might take years for lab research to affect key products, if it ever did. Instead, the focus of our teams was to work on data applications that would have an immediate and massive impact on the business. The term that seemed to fit best was data scientist: those who use both data and science to create something new.

In September 2012, Tom Davenport and D.J. Patil published, "Data Scientist: The Sexiest Job of the 21st Century" in the *Harvard Business Review* An earlier version of this timeline was published in WhatsTheBigData.com.

So, this is how Data Science grew up and became a buzz word for the modern IT Industries.

4. What is the purpose and importance of Data Science

The principal purpose of Data Science is to find patterns within data. It uses various statistical techniques to analyze and draw insights from the data. From data extraction, wrangling and pre-processing, a Data Scientist must scrutinize the data thoroughly. Then, he has the responsibility of making predictions from the data. The goal of a Data Scientist is to derive conclusions from the data. Through these conclusions, he is able to assist companies in making smarter business decisions (https://www.oreilly.com/library/view/doing-data-science/9781449363871/ch01.html[origins of data science]; and https://towardsdatascience.com/how-data-science-will-evolve-over-the-next-decade-ab14fb9c9a75 [origins of data science]).

4.1. Why data is important

Data is the new electricity. We are living in the *age of the fourth industrial revolution*. This is the era of Artificial Intelligence and Big Data. There is a massive data explosion that has resulted in the culmination of new technologies and smarter products. Around *2.5 exabytes of Data* is created each day. The need for data has risen tremendously in the last decade. Many companies have centered their business on data. Today data has created new sectors in IT industry. The questions arise now are (https://towardsdatascience.com/how-data-science-will-evolve-over-the-next-decade-ab14fb9c9a75 [origins of data science]):

1. Why do we need Data?
2. Why do industries need Data?
3. What makes data a precious commodity?

The answers to these questions lies in the way companies have sought to transform.

Data Science is a very recent terminology. Before Data Science, we had statisticians. These statisticians experienced in qualitative analysis of data and companies employed them to analyze their overall performance and sales. With the advent of a computing process, cloud storage, and analytical tools, the field of computer science merged with statistics. This gave birth to Data Science.

Early data analytics based on surveying and finding solutions to public problems. For example, a survey regarding a number of children in a district would lead to a decision of development of the school in that area. With the help of

computers, the decision-making process has been simplified. As a result, computers could solve more complex statistical problems. As Data started to proliferate, companies started to realize its value. Its importance reflected in the many products designed to boost customer experiences. Industries sought experts who could tap the potential that data holstered. Data could help them make the right business decisions and maximize their profits. Moreover, it gave the company an opportunity to examine and act according to customer behavior based on their purchasing patterns. Data helped companies boost their revenue model and helped them craft a better quality product for clients.

Data for products is like electricity for the household gadgets. We need data to engineer the products that cater to the users. It is what drives the product and makes it usable. A data scientist is like a sculptor. He chisels the data to create something meaningful out of it. While it can be a tedious task, a data scientist needs to have the right expertise to deliver the results.

4.2. Why is Data Science important?

Data creates magic. Industries need data to help them make careful decisions. Data Science churns raw data into meaningful insights. Therefore, industries need data science. A Data Scientist is a wizard who knows how to create magic using data. A skilled Data Scientist will know how to dig out meaningful information with whatever data he comes across. He helps the company in the right direction. The company requires strong data-driven decisions at which he's an expert. The Data Scientist is an expert in various underlying fields of Statistics and Computer Science. He uses his analytical aptitude to solve business problems.

While the role of Data Science focuses on the analysis and management of data, it is dependent on the area that the company is specialized in. This requires the Data Scientist to have domain knowledge of that particular industry.

4.3. What is the purpose of Data Centric industries?

Companies need data. They need it for their data-driven decision models and creating better customer experiences. So, let's explore the specific areas that these companies focus on in order to make smarter data-driven decisions (Scott Allen Mongeau, (2008).

I. Data Science for better marketing:

   Companies used Data to analyze their marketing strategies and create better advertisements. Many times, business spends an astronomical amount on marketing their products. This may at times not yield expected results. Therefore, by studying and analyzing customer feedback, companies are able to create better advertisements. The companies do so by carefully analyzing customer behavior online. Also, monitoring customer trends helps the company to get the better market insights. Therefore, business needs Data Scientists to assist them in making strong decisions with regard to marketing campaigns and advertisements.

II. Data Science for customer acquisition:

   Data Scientists help the company to acquire customers by analyzing their needs. This allows the companies to tailor products best suited for the requirements of their potential customers. Data holds the key for companies to understand their clients. Therefore, the purpose of a Data Scientist here is to enable companies to recognize clients and help them deliver the needs of their customers.

III. Data Science for innovation:

   Companies create better innovations with an abundance of data. The Data Scientists aid in product innovation by analyzing and creating insights within the conventional designs. They analyze customer reviews and help the companies craft product that sits perfectly with the reviews and feedback. Using the data from customer feedback, companies make decisions and take proper action in the right direction.

IV. Data Science for enriching lives:

   Customer data is key to making their lives better. Healthcare industries use the data available to them to assist their customers in their everyday life. Data Scientists in these types of industries have the purpose of analyzing the personal data, health history and create products that tackle the problems faced by customers.

   From the above instances of data-centric companies, it is clear that each company uses data differently. The use of data varies as per company requirements. Therefore, the purpose of Data Scientists depends on the interests of the company.

   So, these are the important aspects of Data Science and its purpose. As we can see Data Science is an essential part of industries today and today most of the companies' focus on the data and its types. That's the reason why most of the companies are data centric today.

5. Data Science as a tool for cloud security

We have already seen how much useful data science is for our IT industries apart from that it is also useful for daily analysis like weather forecast, Cricket match prediction etc. Now, the question arises here is,

1. Can data science provide security for the Data?
2. What kind of security is provided by clouds till now?
3. How data science is used as a tool for cloud security?
4. At what level it fails to secure our data?
5. Which kind of penetration could be prevented by the use of Data Science?

Data Science plays an important role in providing security in cloud applications, modern cloud has one most prominent feature that its ubiquitous nature, but it still facing the scalable challenges. In industries that are facing the same problem which is scalability, data science has proven itself highly successful, for example; web search, high-speed finance, high-volume image and video processing and even in large scale defense systems. Recently, data science techniques have adopted increasingly in computer and network security applications. There is no doubt that Data Science can be used as core technology to provide security and to strengthen cloud applications by implementing algorithms which can detect threats in large scale data mining system (Anderson *et al*., 2009).

5.1. The keystone of security is visibility

We can use the Data Science to identify and extract the critical information from the variety of structure and unstructured data by using the techniques such as data mining, machine learning and NLP. The information extracted from the data marts can be used for analysis and to gain insights into the target environment from which data are fetched (Scott Allen Mongeau, 2008).

The keystone of security is visibility. For effective cloud application security, visibility means everything and that means total transparency and understanding. Once this visibility is achieved we can easily detect the malicious files, malware threats, and so that we can easily protect our files and assets having the security systems that interoperate with cloud applications and facilitates alerting, automatic prevention and remediation policies. Data Science plays a significant role in attaining that visibility.

For cloud applications it is a challenge to detect the abnormal user activities, the hacking attempts and the threats that potentially expose or destroy information stored on the cloud service. This necessitates a meaningful level of visibility that captures both user actions and the resources they access.

Conventional security solutions are not designed for cloud applications, the protection they afford to on-premises systems does not effectively translate to the cloud. As service providers continue to simplify these features, the threat of data exfiltration (intentional or accidental) increases, making data loss prevention (DLP) an essential feature of any cloud security solution (Scott Allen Mongeau, 2008).

Lets understand this with an example, an advanced on—premises DLP system does not understand link semantics, so it may not recognize that a link sent over email is associated with a file that breaks Payment Card Industry compliance. The cause can be as simple as the link that is not recognized by DLP system or which simply cannot be accessed or document cannot be easily interpreted by DLP system, or traffic which cannot be interpreted from the site.

Now the question arises here is, can we used Data Science as:

- A tool to ensure a user does not accidently expose a file containing compliance concerns.
- A tool to prevent and remediate data exposure.
- A tool to detect and prevent against a malicious insider, attacker or malware posing as an insider.

The data science can provide all the above listed requirements and concerns. So, let's see how cloud security uses data science to provide consistent and broad visibility into cloud application usage, and detect the new dynamic cloud threats and detect accurately the sensitive content on a cloud service.

5.2. How to achieve visibility in cloud service using Data Science:

In the cloud services the real time visibility is related to protection required during the parsing of HTTP traffic to determine following things:

- The user account accessing the service.

- The actions carried out by the user.
- The resources (e.g., files) accessed or modified.

Visibility in conventional network security is achieved using the static signatures. On the other hand the cloud applications changes its network traffic pattern frequently, that make the manual signature development strained. And by the time it evolved the security patterns so do the challenges to secure the data of the users using the cloud services. This necessitates an approach to signature generation that adapt as quickly as application evolve, while simultaneously scaling the wide breadth of applications as available to its users. Signatures of these cloud applications are typically manual that makes it a time consuming process to build it by hand and as the consumer size increases it is difficult to always change the signature. A time consuming process that is made even more difficult by cloud applications that machine encode critical information such as file names. This is problematic as cloud applications change their traffic patterns, signatures break and it is costly to rebuild them (Sikos and Choo, 2020).

Another problem is numerous applications based on clouds available to users and they all need their individual signature. That results as frequent lack of visibility into how applications are used (Sikos and Choo, 2020; Tariq Mahmood and Uzma Afzal, 2013).

Data science methods can be used to meet this challenge by automatically learning signatures that achieve a zero false positive rate in a fraction of time required for manual construction. As signature breaks these methods can operate within a feedback loop to automatically repair signatures, restoring visibility in a short time.

5.3. Dynamic threat detection

The threats to cloud applications from malicious insiders, attackers and naive users are increasing very frequently. Cloud applications can now be used as host to deliver malware, to establish communication channels for data exfiltration, trigger acts of data destruction, expose critical information and hijack accounts. Specific data science algorithms are in a strong position to provide high quality threat detection when visibility is both rich and meaningful. These algorithms are designed to handle large-scale data analysis and thereby extract meaningful information out of the data. Data science can be used as a tool to detect security issues residing in the cloud because intelligence can be gained on multiple fronts which are:

1) **Correlation –** Mapping large datasets under specific security analytics helps to determine correlation to understand the complete posture of an attack. When data from multiple locations are correlated, attacks can be dissected at a granular level.
2) **Visibility –** Data mining means mining of big datasets and that means big picture visibility. When large data sets are mined, it becomes easier to obtain visibility into the attacks, which at last results in gaining more intelligence.
3) **Baseline –** When the datasets are mined using the specific features related to an attack, it helps to generate baselines that can be used to measure the intensity or amplification of an attack in a given environment.
4) **Context –** Mining of big datasets may provide more adaptive intelligence, including the contextual awareness and situational awareness of a specific attack in the environment.

Data Science algorithms can also effectively integrate multiple information sources to provide a more complete picture of a user's estimated risk to an organization. Such algorithms automatically scale horizontally as the number of input signals increases. Meaningful visibility that logs user action allows for meaningful threat detection (Tariq Mahmood and Uzma Afzal, 2013).

6. Data Science in cybersecurity

Modern Data Science, in its most fundamental form, is all about understanding. It involves studying, processing, and extracting valuable insights from a set of information. Though the word and process have been around for several decades, it was primarily a subset of computer science. Today, it has developed into an independent field and hence, those interested can study and major in it (Sikos and Choo, 2020).

One modern application of Data Science includes cyber security. It may sound strange to study Data Science with the hopes of improving cyber security, but in reality, it makes a lot of sense, and here's why.

6.1. What is relation between Big Data and Cyber Security?

With the help of the Big Data Analytics applications and several machine learning tools and algorithms, it is possible for an organization to conduct analysis of collection of information. Professionals can examine data closely to forecast and reveal the upcoming trends, patterns and actionable Intel.

Lets understand this with an example, let's say one might find out that most of the cyber attacks in their organization takes place in night when no one is in the office. Or maybe one can learn to narrow the probability down to certain terminal on their network. Besides that, one can also use the extracted information to predict potential attacks that can take place in future (Sikos and Choo, 2020; Tariq Mahmood and Uzma Afzal, 2013).

6.2. Predictive and active intrusion detection system

Hackers and attackers use a wide range of tools and ways of intrusion to gain access to the critical information. Now, we use IDS for active and preventive variety.

Modern Data Science can both enhance and simplify the use of such tools. By feeding present and historical data into a Machine Learning algorithm, this system can precisely detect potential problems. Over time, as such system becomes more precise; it can foretell future attacks and spot various loopholes (Tariq Mahmood and Uzma Afzal,2013).

6.3. Protecting the valuable information

Another concern of a data attack is the loss of extremely valuable data and information, which can be really damaging to your organization. With the use of security measures like highly complex signatures or encryption, you can stop anyone from probing into a dataset. By involving Data Science, you can start to build impenetrable protocols. For example, by analyzing the history of your cyber-attacks, you can develop algorithms to detect the most frequently targeted chunks of data.

Overall, Data Science through effective Analytics Systems helps enhance the cyber security industry. It allows IT professionals to come up with more operative, defensive and active measures to prevent cyber-attacks (Tariq Mahmood and Uzma Afzal, 2013).

7. Data Science answer to all cybersecurity challenges

Data science and machine learning continue to improve and advance. One of the areas where it is becoming more relevant is data security – AI in cybersecurity is expected to reach almost $35 bn by 2025.

Data scientists can apply their knowledge to the cyber-security field to help protect against attacks and identify suspicious behavior. The fact that they play a versatile role of a technical expert, problem gatherer, analyst and a skilled interpreter, problem-solving is easiest for them. By using knowledge of data science, coders and programmers can also improve their techniques to create better programs to protect against cyber threats.

In addition, they get benefitted from following thing;

1. Cyber-security industry is always looking for technical resources but they need smart and sharp people who can help them solve problems in no time. So if you're good at coding, you have a high chance of getting hired at sky-high package.
2. Each year, billions are lost in data breaches.

7.1. How data science can be used in cybersecurity programs

In cybersecurity, our goal is to identify threats, stop intrusions and attacks, properly identify malware and spam, and prevent fraud. Data science and machine learning can be used to help better identify these threats. For example, when it comes to identifying malware and spam, data from a wide range of samples can be used for deep learning and training purposes so that malware and spam are properly detected.

The goal here is to properly identify and warn when malware and spam are detected while reducing false positives, which use up unnecessary time and energy. The same goes for identifying intrusions and attacks. When hackers want to attack a system, there will usually be smaller intrusions at first with the intent of figuring out how the system works, what its defenses are, and how they can be overcome. This is commonly the case with Ransomware, of which cases have increased by 37% last year.

Data science can be used to properly identify anomalies and abnormalities in user behavior that may be caused by an intruder. Then, the proper preventative measures can be taken to stop the intrusion from getting more severe. There will often be a correlation of multiple abnormal events if an intrusion or attack is being carried out. Data science can help connect the dots between these "minor" abnormalities and use them to paint a bigger picture of what might be going on. For preventing fraud, the process is the same. Using samples from your data set, you would detect abnormalities in credit card purchases, for example, and use that information to identify fraudulent activity (Tariq Mahmood and Uzma Afzal, 2013; http://www.covert.io/security-datascience-papers/ [Machine Learning tasks in Cyber Security]).

7.2. Challenges to overcome

1. Not relying on "lab-based" sequences:

   One of the main benefits of using data science for cyber-security purposes is that larger samples of data can be used to better identify threats. For example, a common problem with cybersecurity programs is that they were built by using a pre-defined sequence of events.

   However, hackers rarely play by the "rules". It is extremely important to assess all of the real data you have from real users when creating a program to identify threats so that proper normal behavior can be identified, which is essential if you want to identify abnormal behavior.

2. Having access to enough data:

   Identifying malware and spam is a lot easier than identifying behavioral abnormalities. There is a large sample of data available to use for training purposes to identify what is malware and spam and what is not (Scott Allen Mongeau, 2008; Sikos and Choo, 2020). According to Kaspersky, over 360,000 new samples of malware are detected every day.

   On the other hand, when it comes to behavioral abnormalities. There is a lot more nuance involved. We need to assess all the real data that you possibly can in order to know what is normal and what is not instead of relying on preconceived rules.

   As mentioned, data science can be used to assess all raw user behavior and connect the dots if multiple abnormalities are detected. By using large "data lakes," you can compare real-time activity to the data in the lake to help identify threats. The challenge would be having access to all that data, which comes from many different logs and systems.

3. Focusing on the abnormalities

   Not every behavior that is slightly unusual is going to be relevant for cyber-security purposes. Knowledge of why a behavior may have occurred is necessary in order to reduce false positives.

   There are always going to be deviations from so-called normal activity; for example, many people may be traveling to a different country and logging in from there, using a different device to log in, or suddenly deciding to make a purchasing decision that deviates from their previous purchase history.

   A lot depends on the context as well, as the same type of behavior can mean different things depending on what is going on in the bigger picture. There can be a lot of extra noise that is not relevant, thus creating many false positives.

4. Using Data Science in an effective way

   At an enterprise level, data science analyzes big data from the network to root out possible vulnerabilities. On the other hand, data security software, like VPN services, protects the network where the big data flows from. Thus, data science and data security have a symbiotic relationship. On a larger scale, data science can be used to identify trends and movements of malware over time so that impending threats can be anticipated.

   Other ways data science can be used, for example, is by creating a baseline for each user and comparing it with real-time data. Another suggestion is that clustering can be used to create clusters of activities and behaviors which can be classified as abnormal. Data science can be used to reduce false positives and better streamline the alert process so that there isn't an overload of alerts. If the responses to alerts are properly automated, alerts can be properly attended to in real time and the load on security teams can be lessened.

   The fewer false positives, the better. False positives cost companies an average of $1.37 mn a year. Once a program or model is created, it will have to be continually monitored to make sure that it is working as planned. If it is indeed working as planned, the results of the model must be monitored to make sure that they are satisfactory. When combining data science with cyber-security, effort must be taken not to rush into things. You don't want to potentially miss out on attacks and abnormalities due to a lack of proper training.

   In addition, using more than one algorithm can provide better protection in case one algorithm is attacked or corrupted.

8. The Cyber Security Data Science (CSDS) and its importance

From the previous articles we all get aware that data science can provide us security, and it can also trigger the alarm for the upcoming attacks if the security model is trained with proper datasets and training sets. The training model if inserted properly can trace any kind of attack because the cyber attacks done by attackers are usually signature based

and the malwares and other spywares leave signature behind. So, if the signature based malware attacks on the system in an IT industry, the Machine Learning model with application of Data Science can predict the attack and if the attack has happen the Data Science algorithm not only can detect it but also it can even prevent the system from that attack.

But if the attack has happened with the some AI-based malware in that case Data Science fails to detect the attack because the there is no such data which is used to train the system to detect AI- driven cyber attack, now the question raised here is, why Data Science lags to detect the system from such attacks?

It is because right now we don't have enough data to train the model or algorithm in such a way that they can detect an AI-Based malware.

8.1. Why we should worry about AI-Driven Cyber attacks?

The traditional malware can be traced and system can get prevented because they leave signature behind and if malware enters the system the indicators of compromise (IoCs) for it are collected and distributed to antivirus engines. As each file enters a network or computer, it's scanned against the signature list and quarantined or deleted in the event of a match.

The issues with traditional approaches to antivirus detection are delays, scalability and applicability of signatures. By definition, there is always a delay between an attack starting and signatures being available, since someone needs to detect the malware, generate a signature and deploy it. The use of large list of signatures is also problematic since the list grows continuously, making storage and scanning less efficient. Antivirus systems using AI focus on detecting unusual behavior by programs rather than matching signatures. Since most malware is designed to do things that are different from the standard operation of the computer, they can be detected based on these actions. This allows these AI-based AVs to detect zero-day exploits and other previously unknown malware.

But if we talk about AI-driven cyber attacks there is a little bit change of scenario;

Since our AI-system is getting smart enough to differentiate between a normal glitch and malicious attack that could help us to prevent the systems from the normal attacks but if we talk about AI- driven attacks it could get worse because we have a lack of datasets to train our system to differentiate between normal malware attack and AI-driven cyber-attacks. Now here comes the CSDS (http://www.covert.io/security-datascience-papers/ [Machine Learning tasks in Cyber Security]).

8.2. What is cybersecurity data?

We already know data science is largely driven by the availability of data. The one efficient way to trace the cyber-attacks is to take the help of CSDS, in this branch of Data Science we use the cybersecurity data to train our model for searching and detecting cyber attacks. So, the cyber security data is a collection of raw data which contains various types of cyber attacks and their relevant features. The reason of collecting the raw security data is to analyze the various patterns of security incidents or malicious behavior and to build a data-driven security model to achieve the desirable security for our data. Several datasets exist in the area of cybersecurity including intrusion analysis, malware analysis, anomaly, fraud, or spam analysis that is used for various purposes.

8.3. Definition of cybersecurity Data Science

Data Science is transforming the world, and today most of the industries are data-driven. So, it is critically important for the future of intelligent cybersecurity system and services because "security is all about data". When we seek to detect cyber threats, we are analyzing the security data in the form of files, logs, network packets, or other relevant sources. Traditionally, security professionals didn't use data science techniques to make detections based on these data sources. Instead, they used file hashes, custom-written rules like signatures, or manually defined heuristics (Sikos and Choo, 2020).

In recent days, the entire security industry is moving towards data science, because of its capability to transform raw data into decision making. To do this, several data-driven tasks can be associated, such as;

i. Data engineering focusing practical applications of data gathering and analysis.
ii. Reducing data volume that deals with filtering significant and relevant data to further analysis.
iii. Discovery and detection that focuses on extracting insight or incident patterns or knowledge from data.
iv. Automated models that focus on building data-driven intelligent security model.
v. Targeted security alerts focusing on the generation of remarkable security alerts based on discovered knowledge that minimizes the false alerts.

vi. Resource optimization that deals with the avail-able resources to achieve the target goals in a security system. While making data-driven decisions, behavioral analysis could also play a significant role in the domain of cyber security.

Thus, the concept of CSDS incorporates the methods and techniques of data science and machine learning as well as the behavioral analytics of various security incidents. The combination of these technologies has given birth to the term "CSDS", which refers to collect a large amount of security event data from different sources and analyze it using machine learning technologies for detecting security risks or attacks either through the discovery of useful insights or the latest data-driven patterns. It is, however, worth remembering that CSDS is not just about a collection of machine learning algorithms, rather, a process that can help security professionals or analysts to scale and automate their security activities in a smart way and in a timely manner. Therefore, the formal definition can be as follows: "CSDS is a research or working area existing at the intersection of cybersecurity, data science, and machine learning or artificial intelligence, which is mainly security data-focused, applies machine learning methods, attempts to quantify cyber risks or incidents, and promotes inferential techniques to analyze behavioral patterns in security data."

Thus, the concept of CSDS incorporates the methods and techniques of data science and machine learning as well as the behavioral analytics of various security incidents. The combination of these technologies has given birth to the term "CSDS", which refers to collect a large amount of security event data from different sources and analyze it using machine learning technologies for detecting security risks or attacks either through the discovery of useful insights or the latest data-driven patterns. It is, however, worth remembering that CSDS is not just about a collection of machine learning algorithms, rather, a process that can help security professionals or analysts to scale and automate their security activities in a smart way and in a timely manner. Therefore, the formal definition can be as follows: "CSDS is a research or working area existing at the intersection of cybersecurity, data science, and machine learning or artificial intelligence, which is mainly security data-focused, applies machine learning methods, attempts to quantify cyber risks or incidents, and promotes inferential techniques to analyze behavioral patterns in security data. It also focuses on generating security response alerts, and eventually seeks for optimizing cybersecurity solutions, to build automated and intelligent cybersecurity systems."

8.4. Machine learning tasks in cybersecurity

Machine learning (ML) is typically considered as a branch of "Artificial Intelligence", which is closely related to computational statistics, data mining and analytics, data science, particularly focusing on making the computers to learn from data. Thus, machine learning models typically comprise of a set of rules, methods, or complex "transfer functions" that can be applied to find interesting data patterns, or to recognize or predict behavior, which could play an important role in the area of cybersecurity. In the following, we discuss different methods that can be used to solve machine learning tasks and how they are related to cybersecurity tasks (http://www.covert.io/security-datascience-papers/ [Machine Learning tasks in Cyber Security]).

8.5. Supervised learning

Supervised learning is performed when specific targets are defined to reach from a certain set of inputs, i.e., task-driven approach. In the area of machine learning, the most popular supervised learning techniques are known as classification and regression methods. These techniques are popular to classify or predict the future for a particular security problem. For instance, to predict denial-of-service attack (yes, no) or to identify different classes of network attacks such as scanning and spoofing, classification techniques can be used in the cybersecurity domain. ZeroR, OneR, Navies Bayes, Decision Tree, K-nearest neighbors, support vector machines, adaptive boosting, and logistic regression are the well-known classification techniques. In addition, recently Sarker *et al.* (2020b) have proposed BehavDT, and IntruDtree classification techniques that are able to effectively build a data-driven predictive model. On the other hand, to predict the continuous or numeric value, e.g., total phishing attacks in a certain period or predicting the network packet parameters, regression techniques are useful. Regression analyses can also be used to detect the root causes of cybercrime and other types of fraud. Linear regression, support vector regression are the popular regression techniques. The main difference between classification and regression is that the output variable in the regression is numerical or continuous, while the predicted output for classification is categorical or discrete. Ensemble learning is an extension of supervised learning while mixing different simple models, e.g., Random Forest learning that generates multiple decision trees to solve a particular security task.

8.6. Unsupervised learning

In unsupervised learning problems, the main task is to find patterns, structures, or knowledge in unlabeled data, i.e., data-driven approach. In the area of cybersecurity, cyber attacks like malware stays hidden in some ways, include

changing their behavior dynamically and autonomously to avoid detection. Clustering techniques, a type of unsupervised learning, can help to uncover the hidden patterns and structures from the datasets, to identify indicators of such sophisticated attacks. Similarly, in identifying anomalies, policy violations, detecting, and eliminating noisy instances in data, clustering techniques can be useful. K-means, K-medoids are the popular partitioning clustering algorithms, and single linkage or complete linkage is the well-known hierarchical clustering algorithms used in various application domains. Moreover, a bottom-up clustering approach proposed by Sarker *et al.* (2020b) can also be used by taking into account the data characteristics.

8.7. Neural network and deep learning

Deep learning is a part of machine learning in the area of artificial intelligence, which is a computational model that is inspired by the biological neural networks in the human brain. Artificial Neural Network (ANN) is frequently used in deep learning and the most popular neural network algorithm is back propagation. It performs learning on a multi-layer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer. The main difference between deep learning and classical machine learning is its performance on the amount of security data increases. Typically deep learning algorithms perform well when the data volumes are large, whereas machine learning algorithms perform comparatively better on small datasets. In our earlier work, Sarker *et al.* (2020b) we have illustrated the effectiveness of these approaches considering contextual datasets. However, deep learning approaches mimic the human brain mechanism to interpret large amount of data or the complex data such as images, sounds and texts. In terms of feature extraction to build models, deep learning reduces the effort of designing a feature extractor for each problem than the classical machine learning techniques. Beside these characteristics, deep learning typically takes a long time to train an algorithm than a machine learning algorithm, however, the test time is exactly the opposite. Thus, deep learning relies more on high-performance machines with GPUs than classical machine-learning algorithms. The most popular deep neural network learning models include multi-layer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN) or long short-term memory (LSTM) network. In recent days, researchers use these deep learning techniques for different purposes such as detecting network intrusions, malware traffic detection and classification, etc. in the domain of cybersecurity (http://www.covert.io/security-datascience-papers/ [Machine Learning tasks in Cyber Security]; and Iftikhar and Alghamdi, 2009).

9. Upcoming challenges in CSDS

There are several research issues and challenges in the area of CSDS to extract insights from relevant data towards data-driven intelligent decision making for cybersecurity solutions. In the following, the summarized challenges are ranging from data collection to decision making (Tariq Mahmood and Uzma Afzal, 2013).

Cybersecurity datasets: Source datasets are the primary component to work in the area of CSDS. Most of the existing datasets are old and might insufficient in terms of understanding the recent behavioral patterns of various cyber-attacks. Although the data can be transformed into a meaningful understanding level after performing several processing tasks, there is still a lack of understanding of the characteristics of recent attacks and their patterns of happening. Thus, further processing or machine learning algorithms may provide a low accuracy rate for making the target decisions. Therefore, establishing a large number of recent datasets for a particular problem domain like cyber risk prediction or intrusion detection is needed, which could be one of the major challenges in CSDS.

Handling quality problems in cybersecurity datasets: The cyber datasets might be noisy, incomplete, insignificant, imbalanced, or may contain inconsistency instances related to a particular security incident. Such problems in a data set may affect the quality of the learning process and degrade the performance of the machine learning based models. To make a data-driven intelligent decision for cybersecurity solutions, such problems in data is needed to deal effectively before building the cyber models. Therefore, understanding such problems in cyber data and effectively handling such problems using existing algorithms or newly proposed algorithm for a particular problem domain like malware analysis or intrusion detection and prevention is needed, which could be another research issue in CSDS.

Security policy rule generation: Security policy rules reference security zones and enable a user to allow, restrict, and track traffic on the network based on the corresponding user or user group, and service, or the application. The policy rules including the general and more specific rules are compared against the incoming traffic in sequence during the execution, and the rule that matches the traffic is applied. The policy rules used in most of the cybersecurity systems are static and generated by human expertise or ontology-based. Although, association rule learning techniques produce rules from data, however, there is a problem of redundancy generation that makes the policy rule-set complex. Therefore,

understanding such problems in policy rule generation and effectively handling such problems using existing algorithms or newly proposed algorithm for a particular problem domain like access control is needed, which could be another research issue in CSDS.

Context-awareness in cybersecurity: Existing cybersecurity work mainly originates from the relevant cyber data containing several low-level features. When data mining and machine learning techniques are applied to such datasets, a related pattern can be identified that describes it properly. However, a broader contextual information like temporal, spatial, relationship among events or connections, dependency can be used to decide whether there exists a suspicious activity or not. For instance, some approaches may consider individual connections as DoS attacks, while security experts might not treat them as malicious by themselves. Thus, a significant limitation of existing cybersecurity work is the lack of using the contextual information for predicting risks or attacks. Therefore, context-aware adaptive cyber-security solutions could be another research issue in CSDS.

Feature engineering in cybersecurity: The efficiency and effectiveness of a machine learning-based security model has always been a major challenge due to the high volume of network data with a large number of traffic features. The large dimensionality of data has been addressed using several techniques such as principal component analysis (PCA), singular value decomposition (SVD), etc. In addition to low-level features in the datasets, the contextual relationships between suspicious activities might be relevant. Such contextual data can be stored in an ontology or taxonomy for further processing. Thus how to effectively select the optimal features or extract the significant features considering both the low-level features as well as the contextual features, for effective cybersecurity solutions could be another research issue in CSDS.

Remarkable security alert generation and prioritizing: In many cases, the cybersecurity system may not be well defined and may cause a substantial number of false alarms that are unexpected in an intelligent system. For instance, an IDS deployed in a real-world network generates around nine million alerts per day. A network-based intrusion detection system typically looks at the incoming traffic for matching the associated patterns to detect risks, threats or vulnerabilities and generate security alerts. However, to respond to each such alert might not be effective as it consumes relatively huge amounts of time and resources, and consequently may result in self-inflicted DoS. To overcome this problem, a high-level management is required that correlate the security alerts considering the current context and their logical relationship including their prioritization before reporting them to users, which could be another research issue in CSDS (Iftikhar and Alghamdi, 2009; and Norvig and Russell, 2000).

10. Discussion

A literature review to understand cybersecurity data, various defense strategies including intrusion detection techniques, different types of machine learning techniques in cybersecurity tasks. Based on our discussion on existing work, several research issues related to security datasets, data quality problems, policy rule generation, learning methods, data protection, feature engineering, security alert generation, recent analysis etc. are identified that require further research attention in the domain of CSDS.

The scope of CSDS is broad. Several data-driven tasks such as intrusion detection and prevention, access control management, security policy generation, anomaly detection, spam filtering, fraud detection and prevention, various types of malware attack detection and defense strategies, etc. can be considered as the scope of CSDS. Such tasks based categorization could be helpful for security professionals including the researchers and practitioners who are interested in the domain-specific aspects of security systems. The output of CSDS can be used in many application areas such as Internet of things (IoT) security, network security, cloud security, mobile and web applications, and other relevant cyber areas. Moreover, intelligent cybersecurity solutions are important for the banking industry, the healthcare sector, or the public sector, where data breaches typically occur. Besides, the data-driven security solutions could also be effective in AI-based blockchain technology, where AI works with huge volumes of security event data to extract the useful insights using machine learning techniques, and block-chain as a trusted platform to store such data.

Although in this paper, the discussion of CSDS focusing on examining raw security data to data-driven decision making for intelligent security solutions, it could also be related to big data analytics in terms of data processing and decision making. Big data deals with data sets that are too large or complex having characteristics of high data volume, velocity, and variety. Big data analytics mainly has two parts consisting of data management involving data storage, and analytics. The analytics typically describe the process of analyzing such datasets to discover patterns, unknown correlations, rules, and other useful insights. Thus, several advanced data analysis techniques such as AI, data mining, machine learning could play an important role in processing big data by converting big problems to small problems. To do this, the potential strategies like parallelization, divide-and-conquer, incremental learning, sampling, granular computing,

feature or instance selection, can be used to make better decisions, reducing costs, or enabling more efficient processing. In such cases, the concept of CSDS, particularly machine learning-based modeling could be helpful for process automation and decision making for intelligent security solutions. Moreover, researchers could consider modified algorithms or models for handing big data on parallel computing platforms like Hadoop, Storm, etc.

Based on the concept of CSDS discussed in the paper, building a data-driven security model for a particular security problem and relevant empirical evaluation to measure the effectiveness and efficiency of the model, and to assess the usability in the real-world application domain could be a future work.

11. Conclusion

Motivated by the growing significance of cybersecurity and data science, and machine learning technologies, in this paper, I have described how CSDS applies to data-driven intelligent decision making in smart cybersecurity systems and services. We also have discussed how it can impact security data, both in terms of extracting insight of security incidents and the dataset itself. We aimed to work on CSDS by discussing the state of the art concerning security incidents data and corresponding security services. We also discussed how machine learning techniques can impact in the domain of cybersecurity, and examine the security challenges that remain. In terms of existing research, much focus has been provided on traditional security solutions, with less available work in machine learning technique based security systems. For each common technique, we have discussed relevant security research. The purpose of this article is to share an overview of the conceptualization, understanding, modeling, and thinking about CSDS.

We have further identified and discussed various key issues in security analysis to showcase the signpost of future research directions in the domain of CSDS. Based on the knowledge, we have also provided a generic multi-layered frame-work of CSDS model based on machine learning techniques, where the data is being gathered from diverse sources, and the analytics complement the latest data-driven patterns for providing intelligent security services. The framework consists of several main phases like security data collecting, data preparation, machine learning-based security modeling, and incremental learning and dynamism for smart cybersecurity systems and services. We specifically focused on extracting insights from security data, from setting a research design with particular attention to concepts for data-driven intelligent security solutions. Overall, this paper aimed not only to discuss CSDS and relevant methods but also to discuss the applicability towards data-driven intelligent decision making in cybersecurity systems and services from machine learning perspectives. The whole analysis and discussion can have several implications both for security researchers and practitioners. For researchers, we have highlighted several issues and directions for future research. Other areas for potential research include empirical evaluation of the suggested data-driven model, and comparative analysis with other security systems. For practitioners, the multi-layered machine learning-based model can be used as a reference in designing intelligent cybersecurity systems for organizations. I believe that this study on CSDS opens a promising path and can be used as a reference guide for both academia and industry for future research and applications in the area of cybersecurity.

References

Anderson, Ken., Nafus, Dawn., Rattenbury, Tye., and Aipperspach, Ryan. (2009). Numbers Have Qualities Too: Experiences with Ethno Mining. Ethnographic Praxis in Industry Conference Proceedings. 123-140. 10.1111/j.1559-8918.2009.tb00133.x.

Azzah Kabbas, Atheer Alharthi, and Asmaa Munshi (2020). Artificial Intelligence Applications in Cybersecurity. International *Journal of Computer Science and Network Security*. 20(2), 120-124.

Devakunchari, R. and Sourabh, Prakhar Malik (2019). A Study of Cyber Security using Machine Learning Techniques. *International Journal of Innovative Technology and Exploring Engineering*. 8(7C2), 183-186.

Faheem Ullah, Muhammad Ali Babar (2018). Architectural Tactics for Big Data Cybersecurity Analytic Systems: A Review. *Journal of Systems and Software*. 151, 81-118.

http://www.covert.io/security-datascience-papers/ [Machine Learning tasks in Cyber Security].

Iftikhar, B. and Alghamdi, A.S. (2009). Application of artificial neural network within the detection of dos attacks.

Manjeet, Rege and Raymond Blanch K. Mbah (2018). Machine Learning for Cyber Defense and Attack. *The Seventh International Conference on Data Analytics*. Athens, Greece.

Norvig, P. and Russell, S. (2000). *Artificial intelligence: Fashionable approach.A Modern Approach*, Third Edition. Artif. Intell. 175. 935-937. 10.1016/j.artint.2011.01.005.

Sarkar, Iqbal., Kayes, A. S. M., Badsha, Shahriar, Alqahtani, Hamed, Watters, Paul and Ng, Alex. (2020a). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*. 7. 10.1186/s40537-020-00318-5.

Sarker, Iqbal., Colman, Alan., Han, Jun., Asif, Khan., Asif Yoosef., Abushark, B., and Salah, Khaled. (2020b). BehavDT: A Behavioral Decision Tree Learning to Build User-Centric Context-Aware Predictive Model. Mobile Networks and Applications. 25. 10.1007/s11036-019-01443-z.

Sikos, L.F., & Choo, K.K.R. (Eds.). (2020). *Data science in cybersecurity and cyberthreat intelligence*. Springer. https://doi.org/10.1007/978-3-030-38788-4

Sumeet Dua, Xian Du (2011), *Data Mining and Machine Learning in Cybersecurity*. 1st Edition. Auerbach Publications, New York. https://doi.org/10.1201/b10867

Tariq Mahmood and Uzma Afzal (2013). Security Analytics: Big Data Analytics for Cybersecurity. IEEE NCIA Conference at Military College of Signals, Rawalpindi, Pakistan.

Thuraisingham, Bhavani, Khan, Latifur, Masud, Mehedy and Hamlen, Kevin. (2008). Data Mining for Security Applications. 585-589. 10.1109/EUC.2008.62.

Cite this article as: Shiv Hari Tewari (2021). Necessity of data science for enhanced Cybersecurity. *International Journal of Data Science and Big Data Analytics*. 1(1), 63-79. doi: 10.51483/IJDSBDA.1.1.2021.63-79.